PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-157271
(43)Date of publication of application : 30.05.2003

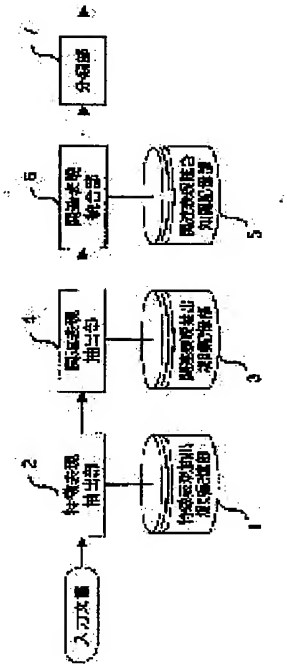
(51)Int.Cl. G06F 17/30

(21)Application number : 2001-355150
(22)Date of filing : 20.11.2001

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : NAGAI AKITO
TAKAYAMA YASUHIRO
SUZUKI KATSUSHI

(54) DEVICE AND METHOD FOR MINING TEXT

(57)Abstract:
PROBLEM TO BE SOLVED: To obtain a text mining device and a method having such a general usability that can analyze even Web pages as analysis objects.
SOLUTION: The text mining device is provided with a feature expression extraction rule storage part 1, a feature expression extraction part 2 for extracting a 1st description unit including intended extraction expression as a feature expression extraction result and outputting the extraction result together with the extraction intention, a related expression extraction rule storage part 3 for allowing a relation name of each 1st description unit to correspond to relation expression, and a related expression extraction part 4 for extracting a 2nd description unit from the feature expression extraction result as a related expression extraction result and outputting the extracted result together with the relation name. The device is provided also with a related expression integrated knowledge storage part 5 for integrating the feature expression and the related expression on the basis of relation of each 1st description unit, a related expression integration part 6 for outputting a related expression integration result and a classification part 7 for classifying the related expression integration result in accordance with the contents of the extraction intention.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－157271
（P2003－157271A）
(43)公開日　平成15年５月30日(2003.5.30)

(51)Int.Cl.[7]	識別記号	ＦＩ	テーマコード（参考）
Ｇ０６Ｆ　17/30	２２０	Ｇ０６Ｆ　17/30	２２０Ｚ　５Ｂ０７５
	１７０		１７０Ａ
	３３０		３３０Ｃ

審査請求　未請求　請求項の数16　ＯＬ　（全 12 頁）

(21)出願番号　特願2001－355150(P2001－355150)

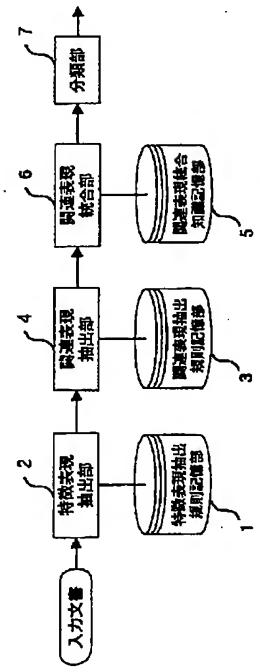
(22)出願日　平成13年11月20日(2001.11.20)

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号
(72)発明者　永井　明人
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(72)発明者　高山　泰博
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(74)代理人　100057874
弁理士　曾我　道照　（外６名）

最終頁に続く

(54)【発明の名称】　テキストマイニング装置および方法

(57)【要約】
【課題】　Ｗｅｂページも含めて分析対象にできるような汎用性を有するテキストマイニング装置および方法を得る。
【解決手段】　特徴表現抽出規則記憶部１と、意図抽出表現を含む第１の記述単位を抽出して特徴表現抽出結果とし、抽出意図とともに出力する特徴表現抽出部２と、第１の記述単位毎の関係名と関係表現とを対応付ける関連表現抽出規則記憶部３と、特徴表現抽出結果から第２の記述単位を抽出して関連表現抽出結果とし、関係名とともに出力する関連表現抽出部４と、特徴表現と関連表現とを第１の記述単位毎の関係に基づいて統合する関連表現統合知識記憶部５と、関連表現統合結果を出力する関連表現統合部５と、関連表現統合結果を抽出意図の内容に応じて分類する分類部７とを備えた。



1

【特許請求の範囲】

【請求項１】　入力文書に関して、抽出意図と意図抽出表現とを対応付けるための特徴表現抽出規則を記憶する特徴表現抽出規則記憶部と、

前記入力文書から、前記意図抽出表現を含む第１の記述単位を抽出して特徴表現とし、前記特徴表現を含む解析結果を特徴表現抽出結果として、前記抽出意図とともに出力する特徴表現抽出部と、

前記第１の記述単位毎の関係に関し、関係名と前記関係を表わす関係表現とを対応付けるための関連表現抽出規則を記憶する関連表現抽出規則記憶部と、

前記特徴表現抽出結果から、前記関係表現を含む第２の記述単位を抽出して関連表現とし、前記関連表現を含む解析結果を関連表現抽出結果として、前記関係名とともに出力する関連表現抽出部と、

前記特徴表現と前記関連表現とを前記第１の記述単位毎の関係に基づいて統合するための関連表現統合知識を記憶する関連表現統合知識記憶部と、

前記関連表現統合知識を参照して前記特徴表現と前記関連表現とを統合し、関連表現統合結果として出力する関連表現統合部と、

関連表現統合結果を前記抽出意図の内容に応じて分類する分類部と

を備えたテキストマイニング装置。

【請求項２】　前記特徴表現抽出部は、前記入力文書を分割して文集合とし、前記意図抽出表現の抽出対象となる第１の記述単位を文とすることを特徴とする請求項１に記載のテキストマイニング装置。

【請求項３】　前記特徴表現抽出部は、前記入力文書における各記述の位置情報を参照し、一定範囲の位置に存在する記述を抽出対象とすることを特徴とする請求項１または請求項２に記載のテキストマイニング装置。

【請求項４】　前記特徴表現抽出部は、文書の構造情報により前記入力文書を分割してサブ文書集合とし、前記サブ文書集合内の各サブ文書毎に前記特徴表現抽出結果を得ることを特徴とする請求項１から請求項３までのいずれかに記載のテキストマイニング装置。

【請求項５】　前記特徴表現抽出部は、文書の構造情報を参照して、前記サブ文書集合の中から、抽出対象とするサブ文書を選択することを特徴とする請求項４に記載のテキストマイニング装置。

【請求項６】　分析の目的とする対象名に関する情報を記憶する対象名辞書と、

前記対象名を含む第３の記述単位を抽出し、前記対象名とともに出力する対象名抽出部とを備え、

前記関連表現統合部は、統合対象となる第１の記述単位である前記特徴表現抽出結果と前記関連表現抽出結果との中から、前記対象名が抽出された第３の記述単位に前記対象名を付与して出力し、

前記分類部は、前記対象名毎に分類して提示することを

2

特徴とする請求項１から請求項５までのいずれかに記載のテキストマイニング装置。

【請求項７】　前記関連表現統合部は、前記特徴表現抽出結果と前記関連表現抽出結果との統合判定に際して、前記対象名を利用することを特徴とする請求項６に記載のテキストマイニング装置。

【請求項８】　前記分類部は、前記関連表現統合部により得られた前記関連表現統合結果を事例として、前記事例のクラスタリングを行い、問題解決木を構成して出力することを特徴とする請求項７に記載のテキストマイニング装置。

【請求項９】　入力文書に関して、抽出意図と意図抽出表現とを対応付けるための特徴表現抽出規則を記憶する特徴表現抽出規則記憶ステップと、

前記入力文書から、前記意図抽出表現を含む第１の記述単位を抽出して特徴表現とし、前記特徴表現を含む解析結果を特徴表現抽出結果として、前記抽出意図とともに出力する特徴表現抽出ステップと、前記第１の記述単位毎の関係に関し、関係名と前記関係を表わす関係表現とを対応付けるための関連表現抽出規則を記憶する関連表現抽出規則記憶ステップと、

前記特徴表現抽出結果から、前記関係表現を含む第２の記述単位を抽出して関連表現とし、前記関連表現を含む解析結果を関連表現抽出結果として、前記関係名とともに出力する関連表現抽出ステップと、

前記特徴表現と前記関連表現とを前記第１の記述単位毎の関係に基づいて統合するための関連表現統合知識を記憶する関連表現統合知識記憶ステップと、

前記関連表現統合知識を参照して前記特徴表現と前記関連表現とを統合し、関連表現統合結果として出力する関連表現統合ステップと、

前記関連表現統合結果を前記抽出意図の内容に応じて分類する分類ステップとを備えたテキストマイニング方法。

【請求項１０】　前記特徴表現抽出ステップは、前記入力文書を分割して文集合とし、前記意図抽出表現の抽出対象となる第１の記述単位を文とすることを特徴とする請求項９に記載のテキストマイニング方法。

【請求項１１】　前記特徴表現抽出ステップは、前記入力文書における各記述の位置情報を参照し、一定範囲の位置に存在する記述を抽出対象とすることを特徴とする請求項９または請求項１０に記載のテキストマイニング方法。

【請求項１２】　前記特徴表現抽出ステップは、文書の構造情報により前記入力文書を分割してサブ文書集合とし、前記サブ文書集合内の各サブ文書毎に前記特徴表現抽出結果を得ることを特徴とする請求項９から請求項１１までのいずれかに記載のテキストマイニング方法。

【請求項１３】　前記特徴表現抽出ステップは、文書の構造情報を参照して、前記サブ文書集合の中から、抽出

対象とするサブ文書を選択することを特徴とする請求項１２に記載のテキストマイニング方法。

【請求項１４】　分析の目的とする対象名に関する情報を記憶する対象名辞書から、前記対象名を含む第３の記述単位を抽出し、前記対象名とともに出力する対象名抽出ステップを備え、

前記関連表現統合ステップは、統合対象となる第１の記述単位である前記特徴表現抽出結果と前記関連表現抽出結果との中から、前記対象名が抽出された第３の記述単位に前記対象名を付与して出力し、

前記分類ステップは、前記対象名毎に分類して提示することを特徴とする請求項９から請求項１３までのいずれかに記載のテキストマイニング方法。

【請求項１５】　前記関連表現統合ステップは、前記特徴表現抽出結果と前記関連表現抽出結果との統合判定に際して、前記対象名を利用することを特徴とする請求項１４に記載のテキストマイニング方法。

【請求項１６】　前記分類ステップは、前記関連表現統合ステップにより得られた前記関連表現統合結果を事例として、前記事例のクラスタリングを行い、問題解決木を構成して出力することを特徴とする請求項１５に記載のテキストマイニング方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、入力される文書中のテキストを解析して特徴的な表現を抽出するテキストマイニング装置および方法に関し、特に、抽出された表現と関係を持つ関連表現をさらに抽出して関係付けることのできるテキストマイニング装置および方法に関するものである。

【０００２】

【従来の技術】近年、インタネット上のＷｅｂページや電子メール、または蓄積されたテキストデータなどの電子化文書は、非常な勢いで増加しており、人手で全文書の内容を読んで調査することは、最早、不可能になってきている。

【０００３】そこで、これらの大量の文書から必要な情報を検索して、内容に応じて分類したり、業務上の分析に役立つ情報を抽出するという技術が、ますます要求されつつある。

【０００４】従来より、上記要求に応える装置として、たとえば特開２００１－１０１１９９号公報（以下、「文献１」という）に記載された「文書処理装置」、または、特開２００１－１４７９３７号公報（以下、「文献２」という）に記載された「業務支援システム」などが提案されている。

【０００５】文献１においては、製品のサポート窓口に送られてきたユーザの電子メールを分析の対象としており、製品の不具合を指摘する電子メールの内容を形態素解析し、形態素を概念に結び付けるための概念分類体系

（シソーラス）を用いて、形態素間の係り受け関係に基づいて、概念間の関係を抽出している。

【０００６】また、文献２においては、営業報告書を分析の対象としており、製品の売れ行き情報に関する概念と表現を定義した概念定義辞書を用いて、営業報告内容から概念を抽出し、接続表現と概念関係との対応を定義した接続関係定義辞書を用いて概念間の関係を抽出している。

【０００７】また、文献２に関連した特開２００１－１８４３５１号公報（以下、「文献３」という）に記載された「文書情報抽出装置および文書分類装置」においては、文献２で用いられる概念定義辞書のメンテナンスを容易にするために、辞書登録すべき候補の一覧を提示するための技術が開示されている。

【０００８】また、特開２０００－３５７１７０号公報（以下、「文献４」という）に記載された「文書の参照理由を用いて情報検索を行う装置」）においては、技術論文のように、文献間に「参照／被参照関係」が存在する場合に、参照理由を同定して、参照理由を用いて文書を検索する装置が開示されている。

【０００９】さらに、立石健二、石黒義英および福島俊一による情報処理学会研究会資料の「インターネットからの評判情報検索」（自然言語処理１４４－１１、第７５～８２頁、２００１年７月１６日）（以下、「文献５」という）においては、インターネットのＷｅｂページから商品の評価表現を抽出する技術が開示されている。

【００１０】図１０は上記文献５に記載されたシステム構成を示すブロック構成図である。図１０において、１０は商品カテゴリに対応した単語を指定する評価表現辞書、１１はクローラ１２および索引ファイル１３を有するＷｅｂページ収集部である。

【００１１】１４は商品名を入力情報とする意見収集部であり、評価表現辞書１０およびＷｅｂページ収集部１１に関連した近接演算処理部１５と、近接演算処理部１５の出力側に挿入された適正値判定処理部１６とを有する。

【００１２】１７は意見収集部１４の出力情報を分析する分類・分析部であり、肯定・否定分類処理部１８を有する。分類・分析部１７による処理情報は、検索結果として出力される。

【００１３】図１０に示した従来システムにおいて、評価表現辞書１０およびＷｅｂページ収集部１１に関連した意見収集部１４と分類・分析部１７とによる評価表現の抽出動作は、以下のようになる。

【００１４】まず、意見収集部１４内の近接演算処理部１５は、Ｗｅｂページ収集部１１に商品名を入力し、索引ファイル１３により得られたインターネット全文検索結果から、別途指定された商品カテゴリに対応する評価表現辞書１０を用いて、評価表現の単語を含む文章を抽

出する。

【００１５】たとえば、商品名が「モバイルギア」であって、商品カテゴリが「コンピュータ」であった場合、評価表現辞書１０に記憶された「速い」、「重い」、「壊れやす」などの単語や、「好き」、「いい」、「勧め」などの単語を含む文章が、全文検索結果から抽出される。

【００１６】次に、分類・分析部１７内の肯定・否定分類処理部１８は、意見収集部１４により抽出された文章に対して、肯定／否定の分類を行い、商品名に対する評価表現の検索結果として出力する。

【００１７】

【発明が解決しようとする課題】従来のテキストマイニング装置および方法は以上のように、文献１および文献２に記載された装置および方法によれば、業務内容に関する概念およびシソーラスを業務内容や製品毎に構築する必要があることから、シソーラスの構築に非常に手間がかかるので、適用する業務内容の変更を容易に行うことができないという問題点があった。

【００１８】また、上記文献１および文献２の装置および方法では、広範囲な概念を含み、且つ新規の概念や新語も発生するＷｅｂページのような文書内容に対しては、あらかじめ概念を定義することができないので、適用することができないという問題点があった。

【００１９】また、文献２に記載された装置および方法によれば、概念間の関係を抽出する処理において、関係を表わす表現（「そこで」、「しかし」など）を文書中より抽出し、これらの表現に対応付けられた関係ＩＤを同定した後に、関係ＩＤの前後に最も近接する概念を関係付けており、概念の関係を隣接する二項関係として扱っているので、抽出すべき概念間の関係を取りこぼしてしまうという問題点があった。

【００２０】たとえば、上記文献２の装置および方法では、「商品Ａは売れている。サンプルを配布したため。しかし、値段が高いという不満がある。」という文書の場合、「値段が高い」という概念は、商品Ａに関する「売れ行きが良い」という概念に対して逆接に関係付けられるべきであるが、最も隣接する「サンプルを配布したため。」に関係付けられてしまうという問題点があった。

【００２１】また、文献３に記載された装置および方法によれば、辞書登録の候補を提示することにより辞書のメンテナンスの容易化を図っているものの、登録の候補は、概念に対応付けられる表現であって、概念自体が固定の定義となっているので、あらかじめ定義した概念の範囲内における辞書メンテナンスのみに限定されるという問題点があった。

【００２２】また、文献４に記載された装置および方法によれば、参照理由の同定に用いてされる「参照特徴－参照理由対応表」において、参照特徴として本文の記述

表現そのものを用いているので、参照理由を表わす表現を全て定義する必要があり、語彙、構文、時制などのバリエーションを考慮すると、参照特徴（定義すべき表現）の数が組合わせ爆発を生じてしまい現実的ではないうえ、参照特徴を抽出する手段（言語解析部）に関して具体的な実現方法の記載がなく、実現性に乏しいという問題点があった。

【００２３】さらに、文献５に記載された装置および方法によれば、抽出対象が評価表現（意見）のみであり、その根拠や理由および原因などの有益な情報が文書中に存在しても抽出することができないという問題点があった。

【００２４】この発明は上記のような問題点を解決するためになされたもので、概念やシソーラスを用いずに表層的な表現に着目して表現抽出を行うことにより、Ｗｅｂページも含めて分析対象にできるような汎用性を有するテキストマイニング装置および方法を得ることを目的とする。

【００２５】また、この発明は、抽出対象となる表現Ａ（意見、クレーム、障害報告など）に加えて、抽出した表現Ａと関係（根拠や理由、原因、例示など）を有する表現Ｂを抽出して、表現Ａ、Ｂを関連付けて提示することにより、有益な情報の発見や、それらの関係の分析を支援することのできるテキストマイニング装置および方法を得ることを目的とする。

【００２６】

【課題を解決するための手段】この発明に係るテキストマイニング装置は、入力文書に関して、抽出意図と意図抽出表現とを対応付けるための特徴表現抽出規則を記憶する特徴表現抽出規則記憶部と、入力文書から、意図抽出表現を含む第１の記述単位を抽出して特徴表現とし、特徴表現を含む解析結果を特徴表現抽出結果として、抽出意図とともに出力する特徴表現抽出部と、第１の記述単位毎の関係に関し、関係名と関係を表わす関係表現とを対応付けるための関連表現抽出規則を記憶する関連表現抽出規則記憶部と、特徴表現抽出結果から、関係表現を含む第２の記述単位を抽出して関連表現とし、関連表現を含む解析結果を関連表現抽出結果として、関係名とともに出力する関連表現抽出部と、特徴表現と関連表現とを第１の記述単位毎の関係に基づいて統合するための関連表現統合知識を記憶する関連表現統合知識記憶部と、関連表現統合知識を参照して特徴表現と関連表現とを統合し、関連表現統合結果として出力する関連表現統合部と、関連表現統合結果を抽出意図の内容に応じて分類する分類部とを備えたものである。

【００２７】また、この発明に係るテキストマイニング装置の特徴表現抽出部は、入力文書を分割して文集合とし、意図抽出表現の抽出対象となる第１の記述単位を文とするものである。

【００２８】また、この発明に係るテキストマイニング

装置の特徴表現抽出部は、入力文書における各記述の位置情報を参照し、一定範囲の位置に存在する記述を抽出対象とするものである。

【００２９】また、この発明に係るテキストマイニング装置の特徴表現抽出部は、文書の構造情報により入力文書を分割してサブ文書集合とし、サブ文書集合内の各サブ文書毎に特徴表現抽出結果を得るものである。

【００３０】また、この発明に係るテキストマイニング装置の特徴表現抽出部は、文書の構造情報を参照して、サブ文書集合の中から、抽出対象とするサブ文書を選択するものである。

【００３１】また、この発明に係るテキストマイニング装置は、分析の目的とする対象名に関する情報を記憶する対象名辞書と、対象名を含む第３の記述単位を抽出し、対象名とともに出力する対象名抽出部とを備え、関連表現統合部は、統合対象となる第１の記述単位である特徴表現抽出結果と関連表現抽出結果との中から、対象名が抽出された第３の記述単位に対象名を付与して出力し、分類部は、対象名毎に分類して提示するものである。

【００３２】また、この発明に係るテキストマイニング装置の関連表現統合部は、特徴表現抽出結果と関連表現抽出結果との統合判定に際して、対象名を利用するものである。

【００３３】また、この発明に係るテキストマイニング装置の分類部は、関連表現統合部により得られた関連表現統合結果を事例として、事例のクラスタリングを行い、問題解決木を構成して出力するものである。

【００３４】また、この発明に係るテキストマイニング方法は、入力文書に関して、抽出意図と意図抽出表現とを対応付けるための特徴表現抽出規則を記憶する特徴表現抽出規則記憶ステップと、入力文書から、意図抽出表現を含む第１の記述単位を抽出して特徴表現とし、特徴表現を含む解析結果を特徴表現抽出結果として、抽出意図とともに出力する特徴表現抽出ステップと、第１の記述単位毎の関係に関し、関係名と関係を表わす関係表現とを対応付けるための関連表現抽出規則を記憶する関連表現抽出規則記憶ステップと、特徴表現抽出結果から、関係表現を含む第２の記述単位を抽出して関連表現とし、関連表現を含む解析結果を関連表現抽出結果として、関係名とともに出力する関連表現抽出ステップと、特徴表現と関連表現とを第１の記述単位毎の関係に基づいて統合するための関連表現統合知識を記憶する関連表現統合知識記憶ステップと、関連表現統合知識を参照して特徴表現と関連表現とを統合し、関連表現統合結果として出力する関連表現統合ステップと、関連表現統合結果を抽出意図の内容に応じて分類する分類ステップとを備えたものである。

【００３５】また、この発明に係るテキストマイニング方法の特徴表現抽出ステップは、入力文書を分割して文集合とし、意図抽出表現の抽出対象となる第１の記述単位を文とするものである。

【００３６】また、この発明に係るテキストマイニング方法の特徴表現抽出ステップは、入力文書における各記述の位置情報を参照し、一定範囲の位置に存在する記述を抽出対象とするものである。

【００３７】また、この発明に係るテキストマイニング方法の特徴表現抽出ステップは、文書の構造情報により入力文書を分割してサブ文書集合とし、サブ文書集合内の各サブ文書毎に特徴表現抽出結果を得るものである。

【００３８】また、この発明に係るテキストマイニング方法の特徴表現抽出ステップは、文書の構造情報を参照して、サブ文書集合の中から、抽出対象とするサブ文書を選択するものである。

【００３９】また、この発明に係るテキストマイニング方法は、分析の目的とする対象名に関する情報を記憶する対象名辞書から、対象名を含む第３の記述単位を抽出し、対象名とともに出力する対象名抽出ステップを備え、関連表現統合ステップは、統合対象となる第１の記述単位である特徴表現抽出結果と関連表現抽出結果との中から、対象名が抽出された第３の記述単位に対象名を付与して出力し、分類ステップは、対象名毎に分類して提示するものである。

【００４０】また、この発明に係るテキストマイニング方法の関連表現統合ステップは、特徴表現抽出結果と関連表現抽出結果との統合判定に際して、対象名を利用するものである。

【００４１】また、この発明に係るテキストマイニング方法の分類ステップは、関連表現統合ステップにより得られた関連表現統合結果を事例として、事例のクラスタリングを行い、問題解決木を構成して出力するものである。

【００４２】

【発明の実施の形態】実施の形態１．以下、図面を参照しながら、この発明の実施の形態１について詳細に説明する。図１はこの発明の実施の形態１によるテキストマイニング装置を示すブロック構成図である。

【００４３】図１において、１は特徴表現抽出規則記憶部であり、入力文書に関して、抽出意図と意図抽出表現とを対応付けるための特徴表現抽出規則を記憶する。２は特徴表現抽出規則記憶部１と関連した特徴表現抽出部であり、取り込まれた入力文書から、意図抽出表現を含む第１の記述単位を抽出して特徴表現とし、特徴表現を含む解析結果を特徴表現抽出結果として抽出意図とともに出力する。

【００４４】３は関連表現抽出規則記憶部であり、第１の記述単位毎の関係に関し、関係名と関係を表わす関係表現とを対応付けるための関連表現抽出規則を記憶する。４は関連表現抽出規則記憶部３と関連した関連表現抽出部であり、特徴表現抽出結果から関係表現を含む第

２の記述単位を抽出して関連表現とし、関連表現を含む解析結果を関連表現抽出結果として関係名とともに出力する。

【００４５】５は関連表現統合知識記憶部であり、関連表現抽出部４から取得される特徴表現と関連表現とを、第１の記述単位毎の関係に基づいて統合するための関連表現統合知識を記憶する。

【００４６】６は関連表現統合知識記憶部５と関連した関連表現統合部であり、関連表現統合知識を参照して特徴表現と関連表現とを統合し、関連表現統合結果として出力する。７は関連表現統合結果を内容に応じて分類する分類部である。

【００４７】次に、図２のフローチャートおよび図３～図８の説明図を参照しながら、図１に示したこの発明の実施の形態１により実行されるテキストマイニング処理動作について説明する。

【００４８】図２において、まず、特徴表現抽出部２に入力文書が取り込まれる（ステップＳ２０１）。なお、入力文書は、たとえば、インターネット上に存在するＷｅｂページや、電子メール、記憶装置や記憶媒体に記録された文書ファイルのデータなどであり、電子化されたテキストであれば何でもよい。

【００４９】続いて、特徴表現抽出部２は、入力文書から記述単位Ｕが取得可能か否かを判定し（ステップＳ２０２）、記述単位Ｕが取得可能（すなわち、ＹＥＳ）と判定されれば、後述の判定ステップＳ２０３に進み、記述単位Ｕが取得不可能（すなわち、ＮＯ）と判定されれば、後述のステップＳ２０６に進む。

【００５０】なお、記述単位Ｕは、入力文書中のテキストの一部（または、全て）であってもよく、特徴表現抽出部２は、たとえば、句点、疑問符、改行などの字句情報から一文単位を切り出して記述単位Ｕとするか、または、入力文書中の単語の位置情報を参照し、一定範囲の位置に存在する記述を切り出して記述単位Ｕとする。

【００５１】または、特徴表現抽出部２は、ＨＴＭＬ文書のような構造を持った文書に対してタグ情報を参照し、たとえばインターネットの掲示板であれば、個別の記事毎に入力文書を分割してサブ文書とする。

【００５２】さらに、特徴表現抽出部２は、タグ情報を参照して、テキスト情報が含まれている記述を選択して記述単位Ｕとする。こうして切り出された記述単位Ｕは、一時的なバッファＴに格納される。

【００５３】ステップＳ２０２において、記述単位Ｕが取得可能（すなわち、ＹＥＳ）と判定された場合、特徴表現抽出部２は、取得した記述単位Ｕに対して、特徴表現抽出規則記憶部１に記憶された抽出意図と意図抽出表現とを参照し、記述単位Ｕに意図抽出表現が含まれているか否かを判定する（ステップＳ２０３）。

【００５４】記述単位Ｕに意図抽出表現が含まれている（すなわち、ＹＥＳ）と判定されれば、後述のステップＳ２０４に進み、記述単位Ｕに意図抽出表現が含まれていない（すなわち、ＮＯ）と判定されれば、ステップＳ２０２に戻る。

【００５５】たとえば、抽出意図が「クレーム」、「要望」、「賞賛」の場合、特徴表現抽出規則記憶部１内の規則は、図３のように定義される。図３において、特徴表現抽出規則は、各抽出意図毎の意図抽出表現として対応付けられ、記述単位Ｕ内の単語の共起パターンが、「１．０」、「０．８」などの重み付きで表現された形式で表される。

【００５６】ステップＳ２０３において、意図抽出表現が記述単位Ｕに含まれる（すなわち、ＹＥＳ）と判定された場合、この記述単位Ｕを特徴表現として抽出し、抽出された記述単位Ｕに抽出意図のラベルを付与する（ステップＳ２０４）。また、記述単位Ｕを特徴表現リストに追加して（ステップＳ２０５）、ステップＳ２０２に戻る。

【００５７】特徴表現リストは、たとえば、図４に示すような情報を持つリストである。図４において、「Ａ社のサポート受付は対応が悪い」など特徴表現に対して、「クレーム」などの抽出意図、および、「２１－３６（文字）」などの「開始－終了」の位置が対応付けられ、さらに、「１．０」などの重みが付けられる。

【００５８】上記特徴表現リストの作成処理（ステップＳ２０２～Ｓ２０５）は、記述単位Ｕが取得できる限り、繰り返し実行される。その後、取得すべき記述単位Ｕが存在しなくなった（ステップＳ２０２でＮＯと判定された）時点で、特徴表現抽出部２は、特徴表現リストの作成処理を終了し、特徴表現リスト（図４参照）と、一時的なバッファＴに格納されている記述単位群（たとえば、図５参照）とを、関連表現抽出部４に出力する。

【００５９】これにより、関連表現抽出部４は、上記処理ステップＳ２０２～Ｓ２０５で得られた特徴表現リストを読み込み（ステップＳ２０６）、各特徴表現に関して、以下の処理を実行する。

【００６０】まず、特徴表現リストが空か否かを判定し（ステップＳ２０７）、特徴表現リストが空である（すなわち、ＹＥＳ）と判定されれば、後述のステップＳ２１２に進む。

【００６１】また、特徴表現リストが空でない（すなわち、ＮＯ）と判定されれば、関連表現抽出部４は、特徴表現の位置情報を参照し、一時的なバッファＴに格納されている記述単位群の中から、特徴表現から一定範囲に存在する記述単位Ｒを取得する（ステップＳ２０８）。

【００６２】たとえば、図５に示すような記述単位群の場合、番号「１」の記述単位が抽出意図「賞賛」の特徴表現として抽出されているものとする。また、一定範囲としては、特徴表現の終了位置から後方「８０文字」の範囲が設定されているものとする。

【００６３】図５に示した記述単位群の場合、番号

「１」の記述単位に関する位置情報（開始－終了）を参照すると「１－２６」であり、一定範囲「８０文字」内に存在する記述単位Ｒは、番号「２」～「４」となる。

【００６４】次に、取得された記述単位Ｒに対して、関連表現抽出部４は、関連情報抽出規則記憶部３に記憶された「関係名」と「関係を表わす関係表現」とを参照し、記述単位Ｒに関係表現が含まれているか否かを判定する（ステップＳ２０９）。

【００６５】ステップＳ２０９において、記述単位Ｒに関係表現が含まれている（すなわち、ＹＥＳ）と判定されれば、後述のステップＳ２１０に進み、記述単位Ｒに関係表現が含まれていない（すなわち、ＮＯ）と判定されれば、ステップＳ２０７に戻る。

【００６６】関連情報抽出規則記憶部３内の規則は、たとえば、図６のように定義されている。図６において、関係表現としては、言語的な意味関係を表わす接続詞や接続助詞（たとえば、「というのは」、「だって」、「（だ）から」）などが関係名毎に定義される。

【００６７】ステップＳ２０９において、関係表現が記述単位Ｒに含まれる（すなわち、ＹＥＳ）と判定された場合、関連表現抽出部４は、この記述単位Ｒを関連表現として抽出する。

【００６８】たとえば、図５において、番号「３」の記述単位Ｒ「すっきりした感じで、コクもあるからです」の中には、図６に参照される関係名「理由」に対応した関係表現「（だ）から」が存在するため、関連表現として番号「３」を抽出することになる。

【００６９】次に、関連表現として抽出された記述単位Ｒに関係名のラベルを付与し（ステップＳ２１０）、記述単位Ｒを関連表現リストに追加して（ステップＳ２１１）、ステップＳ２０７に戻る。

【００７０】関連表現リストは、たとえば、図７に示すような情報を持つリストとして表現される。図７において、たとえば、番号「３」の関連表現「すっきりした感じで、コクもあるからです」に対応して、関係名「理由」および位置「５３－７１」が対応付けられる。

【００７１】上記関連表現リストの作成処理（ステップＳ２０７～Ｓ２１１）は、記述単位Ｒが取得できる限り、繰り返し実行される。その後、取得すべき記述単位Ｒが存在しなくなった（ステップＳ２０７において、ＮＯと判定された）時点で、関連表現抽出部４は、関連表現リストの作成処理を終了し、特徴表現リスト（図４参照）と、関連表現リスト（図７参照）とを、関連表現統合部６に出力する。

【００７２】これにより、関連表現統合部６は、関連表現統合知識記憶部５を参照して特徴表現と関連表現とを統合し（ステップＳ２１２）、関連表現統合結果として分類部７に出力する。

【００７３】ここで、関連表現統合知識記憶部５は、ある特徴表現に統合し得る関連表現を判定するための関連表現統合知識を記憶している。たとえば、以下のような判定基準を用い、各関連表現に対して統合可能性のスコアを算定する。

【００７４】位置情報に関しては、関連表現の開始位置が、特徴表現の終了位置に近いほどスコアを高く算定するか、または、関連表現の終了位置が、特徴表現の開始位置に近いほどスコアを高く算定する。

【００７５】同一単語に関しては、特徴表現と関連表現とで共通の単語が存在すればスコアを加算する。また、類義語に関しては、特徴表現と関連表現とで類義の単語が存在すればスコアを加算する。

【００７６】単語の類義性の判定方法としては、たとえば、特徴表現および関連表現に出現する単語に関して、単語の出現頻度に基づく統計的な重み付けを行って作成した索引表を用いて、文章同士の類似度を単語ベクトルの内積として算定する方法がある。または、類義語辞書を使用する判定方法も可能である。

【００７７】関連表現統合部６は、上記のように算定されたスコアの高い関連表現を特徴表現に統合し、関連表現統合結果を生成する。関連表現統合結果は、たとえば、図８に示すように、特徴表現に対して統合された関連表現を「木構造」の形式で構成したものとなる。

【００７８】図８に示した関連表現統合結果の例では、抽出意図が番号「１」の「賞賛」に対し、番号「３」の関係名「理由」と、番号「４」の関係名「逆接」とが統合されている。

【００７９】図２に戻り、最後に、分類部７は、関連表現統合部６から入力された関連表現統合結果を内容に応じて分類し（ステップＳ２１３）、図２の処理を終了する。このとき、分類部７は、たとえば、関連表現統合結果に付与されている抽出意図（図３、図４、図８参照）に応じて分類を行う。

【００８０】上記処理手順により、分析対象に依存する概念定義やシソーラスを用いずに、表層的な表現に着目して表現抽出を行うことができ、分析対象に依存しない汎用性を有しつつ、一般のＷｅｂページも含めて分析することができる。

【００８１】また、抽出対象となる表現Ａ（クレーム、要望、賞賛などの抽出意図）に加えて、抽出した表現Ａに対して関係（理由、例示、逆接などの関係名）を有する表現Ｂを抽出し、表現Ａ、Ｂを関連付けて提示することにより、有益な情報の発見や、それらの関係の分析を支援することができる。

【００８２】実施の形態２．なお、上記実施の形態１では、関連表現統合知識記憶部５内の知識のみを関連表現統合部６と関連させたが、さらに対象名辞書内の情報を関連させてもよい。

【００８３】図９はこの発明の実施の形態２によるテキストマイニング装置を示すブロック構成図であり、前述（図１参照）と同様のものについては、同一符号を付し

て詳述を省略する。

【００８４】８は対象名辞書であり、分析の目的とする対象名に関する情報を記憶する。９は対象名抽出部であり、対象名辞書８と関連表現統合部６との間に介在されており、対象名を含む第３の記述単位を抽出して、対象名とともに出力する。

【００８５】この場合、関連表現統合部６は、統合対象となる第１の記述単位である特徴表現抽出結果（図４に参照の特徴表現リスト）と関連表現抽出結果（図７に参照の関連表現リスト）との中から、対象名が抽出された第３の記述単位に対象名を付与して分類部７に出力する。

【００８６】これにより、分類部７は、関連表現統合部６から入力された関連表現統合結果に付与されている対象名毎に応じて、関連表現統合結果を分類して提示する。

【００８７】また、関連表現統合部６は、特徴表現抽出結果と関連表現抽出結果との統合判定に際して、対象名を利用する。さらに、分類部７は、関連表現統合部６により得られた関連表現統合結果を事例として、事例のクラスタリングを行い、問題解決木を構成して出力する。

【００８８】図９において、図１と異なる点は、対象名辞書８および対象名抽出部９を備えていることのみである。対象名辞書８に記憶される対象名としては、たとえば、企業名、製品名、機器名、部品名、人名などの固有名詞である。

【００８９】対象名辞書８は、上記の固有名詞をリストとして記憶し、対象名抽出部９は、与えられた記述単位中から対象名を抽出し、該当する記述単位の情報として対象名を付与する。

【００９０】したがって、関連表現統合部６は、ある特徴表現に統合し得る関連表現を判定するための知識として、対象名抽出部９で付与された対象名を利用することにより、関連表現が含まれない記述単位Ｒに関しても、対象名に関する記述を有する関連情報として、特徴表現に統合することができる。

【００９１】また、分類部７は、特徴表現抽出部２により抽出された特徴表現と、関連表現抽出部４により抽出された関連表現（理由や例示などの関係を有する）と、関連表現統合部６により統合された対象名とからなる組を事例として、事例のクラスタリングを行い、「問題解決木」を構成して出力することができる。なお、クラスタリング技術については、公知の方法が種々提案されており、ここでは、どの方法を用いても構わない。

【００９２】上記のように、図９の構成を用いて処理することにより、分析の目的とした対象名に関する、競合製品や他社の評判、関連する機器名／部品名の関連情報を抽出することができる。

【００９３】また、分類部７において、事例のクラスタリングを行うことにより、分析の目的とした対象名に関する、不評、クレーム、障害発生などの問題に関して、共通の現象（特徴表現）に関する理由や具体例などの分析を効果的に支援することができる。

【００９４】

【発明の効果】以上のように、この発明によれば、入力文書に関して、抽出意図と意図抽出表現とを対応付けるための特徴表現抽出規則を記憶する特徴表現抽出規則記憶部と、入力文書から、意図抽出表現を含む第１の記述単位を抽出して特徴表現とし、特徴表現を含む解析結果を特徴表現抽出結果として、抽出意図とともに出力する特徴表現抽出部と、第１の記述単位毎の関係に関し、関係名と関係を表わす関係表現とを対応付けるための関連表現抽出規則を記憶する関連表現抽出規則記憶部と、特徴表現抽出結果から、関係表現を含む第２の記述単位を抽出して関連表現とし、関連表現を含む解析結果を関連表現抽出結果として、関係名とともに出力する関連表現抽出部と、特徴表現と関連表現とを第１の記述単位毎の関係に基づいて統合するための関連表現統合知識を記憶する関連表現統合知識記憶部と、関連表現統合知識を参照して特徴表現と関連表現とを統合し、関連表現統合結果として出力する関連表現統合部と、関連表現統合結果を抽出意図の内容に応じて分類する分類部とを備え、概念やシソーラスを用いずに表層的な表現に着目して表現抽出を行うようにしたので、Ｗｅｂページも含めて分析対象にできるような汎用性を有するテキストマイニング装置が得られる効果がある。また、抽出対象となる表現Ａ（意見、クレーム、障害報告など）に加えて、抽出した表現Ａと関係（根拠や理由、原因、例示など）を有する表現Ｂを抽出して、表現Ａ、Ｂを関連付けて提示するようにしたので、有益な情報の発見や、それらの関係の分析を支援することのできるテキストマイニング装置が得られる効果がある。

【００９５】また、この発明によれば、分析の目的とする対象名に関する情報を記憶する対象名辞書と、対象名を含む第３の記述単位を抽出し、対象名とともに出力する対象名抽出部とを備え、関連表現統合部は、統合対象となる第１の記述単位である特徴表現抽出結果と関連表現抽出結果との中から、対象名が抽出された第３の記述単位に対象名を付与して出力し、分類部は、対象名毎に分類して提示するようにしたので、対象名に関する、競合製品や他社の評判、関連する機器名／部品名の関連情報を抽出することのできるテキストマイニング装置が得られる効果がある。

【００９６】また、この発明によれば、分類部は、関連表現統合部により得られた関連表現統合結果を事例として、事例のクラスタリングを行い、問題解決木を構成して出力するようにしたので、対象名に関する、不評、クレーム、障害発生などの問題に関して、共通の現象（特徴表現）に関する理由や具体例などの分析を効果的に支援することのできるテキストマイニング装置が得られる

効果がある。

【００９７】また、この発明によれば、入力文書に関して、抽出意図と意図抽出表現とを対応付けるための特徴表現抽出規則を記憶する特徴表現抽出規則記憶ステップと、入力文書から、意図抽出表現を含む第１の記述単位を抽出して特徴表現とし、特徴表現を含む解析結果を特徴表現抽出結果として、抽出意図とともに出力する特徴表現抽出ステップと、第１の記述単位毎の関係に関し、関係名と関係を表わす関係表現とを対応付けるための関連表現抽出規則を記憶する関連表現抽出規則記憶ステップと、特徴表現抽出結果から、関係表現を含む第２の記述単位を抽出して関連表現とし、関連表現を含む解析結果を関連表現抽出結果として、関係名とともに出力する関連表現抽出ステップと、特徴表現と関連表現とを第１の記述単位毎の関係に基づいて統合するための関連表現統合知識を記憶する関連表現統合知識記憶ステップと、関連表現統合知識を参照して特徴表現と関連表現とを統合し、関連表現統合結果として出力する関連表現統合ステップと、関連表現統合結果を抽出意図の内容に応じて分類する分類ステップとを備え、概念やシソーラスを用いずに表層的な表現に着目して表現抽出を行うようにしたので、Ｗｅｂページも含めて分析対象にできるような汎用性を有するテキストマイニング方法が得られる効果がある。

【００９８】また、この発明によれば、分析の目的とする対象名に関する情報を記憶する対象名辞書から、対象名を含む第３の記述単位を抽出し、対象名とともに出力する対象名抽出ステップを備え、関連表現統合ステップは、統合対象となる第１の記述単位である特徴表現抽出結果と関連表現抽出結果との中から、対象名が抽出された第３の記述単位に対象名を付与して出力し、分類ステップは、対象名毎に分類して提示するようにしたので、対象名に関する、競合製品や他社の評判、関連する機器名／部品名の関連情報を抽出することのできるテキストマイニング方法が得られる効果がある。

＊【００９９】また、この発明によれば、分類ステップは、関連表現統合ステップにより得られた関連表現統合結果を事例として、事例のクラスタリングを行い、問題解決木を構成して出力するようにしたので、対象名に関する、不評、クレーム、障害発生などの問題に関して、共通の現象（特徴表現）に関する理由や具体例などの分析を効果的に支援することのできるテキストマイニング方法が得られる効果がある。

【図面の簡単な説明】

【図１】　この発明の実施の形態１によるテキストマイニング装置を示すブロック構成図である。

【図２】　この発明の実施の形態１による処理動作を示すフローチャートである。

【図３】　この発明の実施の形態１による特徴表現抽出規則の例を示す説明図である。

【図４】　この発明の実施の形態１による特徴表現リストの例を示す説明図である。

【図５】　この発明の実施の形態１による記述単位群の例を示す説明図である。

【図６】　この発明の実施の形態１による関連表現抽出規則の例を示す説明図である。

【図７】　この発明の実施の形態１による関連表現リストの例を示す説明図である。

【図８】　この発明の実施の形態１による関連表現統合結果の例を示す説明図である。

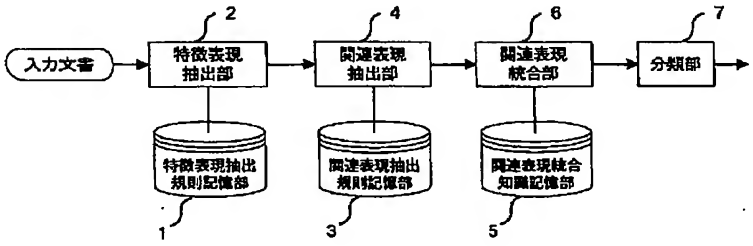
【図９】　この発明の実施の形態２によるテキストマイニング装置を示すブロック構成図である。

【図１０】　従来のテキストマイニング装置を示すブロック構成図である。

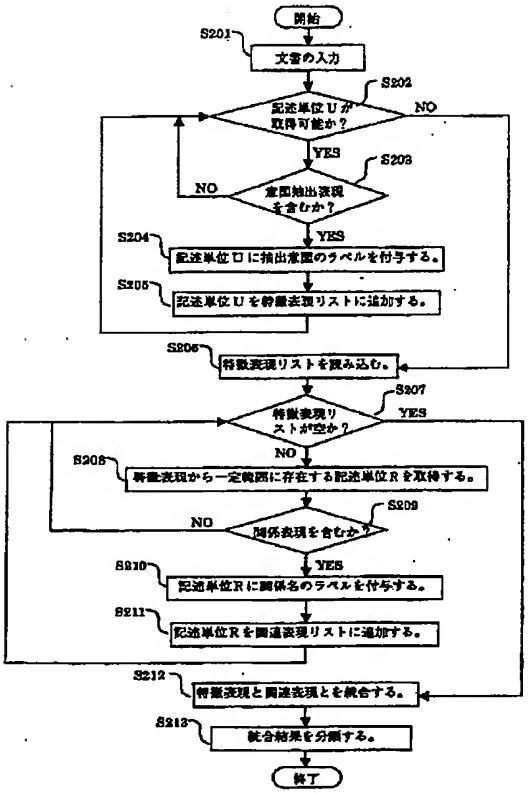
【符号の説明】

１　特徴表現抽出規則記憶部、２　特徴表現抽出部、３　関連表現抽出規則記憶部、４　関連表現抽出部、５　関連表現統合知識記憶部、６　関連表現統合部、７　分類部、８　対象名辞書、９　対象名抽出部。

【図１】



【図２】



【図３】

特徴表現抽出規則の例

抽出意図	意図抽出表現	重み
クレーム	・対応 / 悪 ・納得 / でき / ない	1.0 1.0
要望	・し / て / 欲	0.8
賞賛	・とても / 満足	1.0

【図４】

特徴表現リストの例

特徴表現	抽出意図	位置(開始－終了)	重み
「A社のサポート受付は対応が悪い」	クレーム	21 - 36	1.0
「(製品名)はすぐ壊れるのに、値段が高いのには納得できない」	クレーム	119 - 148	1.0
「講習会を開催して欲しいなあと思います」	要望	99 - 117	0.8
「この秋に出た(製品名)の味にとても満足しています」	賞賛	1 - 26	1.0

【図６】

関連表現抽出規則の例

関係名	関係表現
理由	「というのは」「だって」「(だ)から」「のため」「に起因」
例示	「例えば」「一例」
順接	「それで」「そこで」「つまり」
逆接	「しかし」「でも」「ところが」「(だ) けど」
並列	「また」「さらに」「一方」
…	…

【図５】

記述単位群の例

番号	記述単位群	位置(開始－終了)
1	「この秋に出た(製品名)の味にとても満足しています」	1 - 26
2	「これは今までの製品とはちょっと違うような気がします」	27 - 52
3	「すっきりした感じでコクもあるからです」	53 - 71
4	「でも、デザインがいまいちですね」	72 - 87
5	「まあ、いいか」	88 - 94

【図７】

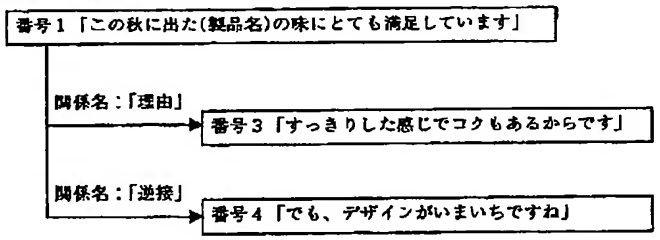
関連表現リストの例

番号	関連表現	関係名	位置(開始－終了)
3	「すっきりした感じで、コクもあるからです」	理由	53 - 71
4	「でも、デザインがいまいちですね」	逆接	72 - 87
…	…	…	…

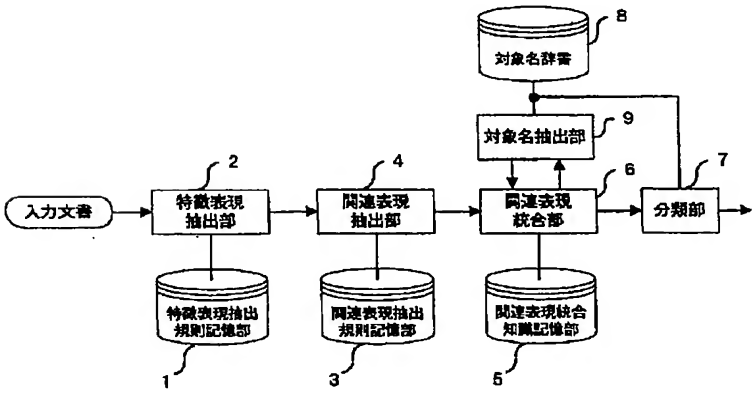
【図８】

関連表現統合結果の例

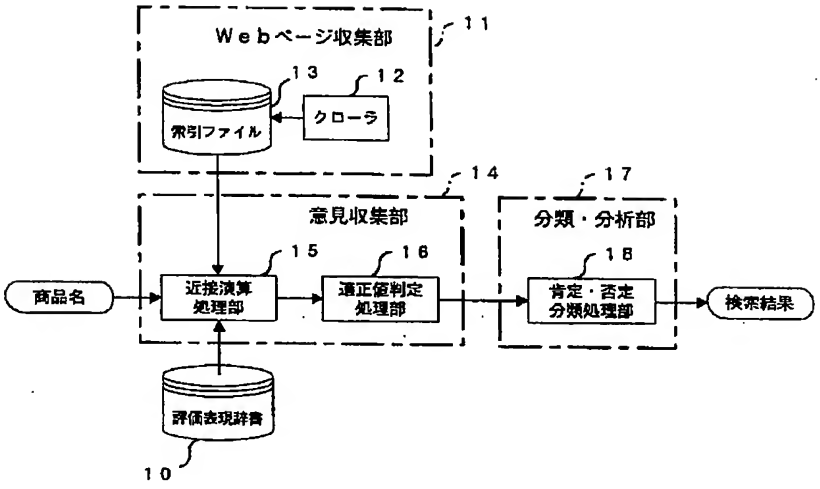
抽出意図：賞賛



【図９】



【図１０】



フロントページの続き

(72)発明者　鈴木　克志
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

Ｆターム(参考)　5B075 ND03 NK06 PP24 UU06